04｜如何评价音频质量的好与坏？

2021-11-29 冯建元

《搞定音频技术》 课程介绍 >

讲述：冯建元

时长 13:38　大小 12.49M

你好，我是建元。

音频在采集、处理、压缩、传输、播放等过程中，不可避免的会对音频的质量产生影响。这节课我们就一起来看一下，音频质量是如何进行评价的。

你是否还记得，在之前我们讲音频编码封装的时候，讲过有损编码和无损编码。对于有损编码，有的人听上去几乎和无损一致，有的人却能听出差别。再比如降噪算法，降噪算法可能会对人声产生损伤、也可能有残余噪声影响听感。因此，评判一个降噪算法的好坏也需要一套综合的音频评价体系。

其实音频的评价方法主要有两种。一种是主观评价，即组织足够数量的人来听被测音频样本，并给每个被测样本打分，最后根据测试人打分的高低来评判音频质量的好坏。**主观测**

试是音频评价的黄金准则，这样的评价是最符合人的实际听感的。但是**主观评测费时费力**，在算法迭代、研发等中间过程中不一定是最经济的方案。

另一种测试方法是客观测试，即通过数学方式计算出一些音频质量评价所需要的指标，比如信噪比（SNR）、频谱差异等。然后综合这些指标去拟合一个主观分数。这样就可以通过数学计算而不是人来给出一个音频质量的评价。

但无论是主观测试还是客观测试，都需要遵守一套严格的测试评价标准，来保证音频质量评价的准确性。那么下面我们先来看看主观评价有哪些可用的方法。

音频主观评价方法

在介绍主观测试方法前，先介绍一个组织：ITU（国际电信联盟）。ITU 是联合国下属的一个专门机构，负责电信、通话等相关标准的制定。其中的无线电通信组（ITU-R）和远程通信标准化组织（ITU-T）为了统一国际的音频质量评价方法制定了一系列的主、客观评价方法。按照 ITU 的评价方法，你的音频算法或者系统的评价结果才会比较有公信力。好了，下面就让我们以 ITU 主观评价方法中的 MUSHRA 为例来介绍一下如何做音频主观评价。

MUSHRA

MUSHRA（Multi-Stimulus Test with Hidden Reference and Anchor，多激励隐藏参考基准测试方法）属于 ITU-R BS.1534 中的推荐测试方法。它最早被用于流媒体与通信的相关编码的主观评价，现在也被广泛应用于心理声学相关研究中的音质主观评价。其测试的特点主要是**在测试语料中混入无损音源作为参考（上限），全损音源作为锚点（下限），通过双盲听测试，对待测音源和隐藏参考音源与锚点进行主观评分。**

所谓**“双盲”就是测试人和提供测试的人都不知道自己要听的是哪段语料**，这个在测试环节中很重要。比如，如果你提前告诉测试人“你的算法会让声音中的风声不那么刺耳”之类的暗示，或者在测试的时候双号为无损语料，单号为测试语料，这样的操作都会让测试结果不具有参考性。

MUSHRA 的分数是 0 到 100 分，按照从高到低的听感描述，如表 1 所示：

分数	80~100分	60~80分	40~60分	20~40分	0~20分
描述	非常好	好	一般	差	非常差

极客时间

表1 MUSHRA分数代表的听感

测试规范中有很多测试细节，这里我罗列一下。为了保证测试的有效性，主要需要注意以下 4 点：

1. 参考的标准音频和被测试音频间隔测试，连续重复 4 次；
2. 音频源采用 15 ~ 20s；
3. 一次完整的测试时间不应超过 15 ~ 20min；
4. 测试成员：专家成员最少 10 人，非专家 20 人。

这里我解释一下为什么要注意这些细节。

首先，测试语料需要重复 4 次主要是为了，防止只出现一次可能会导致误判。比如，听音者一开始没有集中精神打出了偏低的分数。

其次，音频源的时间控制是为了防止听音者产生听音疲劳，从而导致判断失误。

然后，由于集中精神听声音的细节就和瑜伽的正念训练一样，是一件耗费心神的事情，所以不宜时间太长。因此，一般测试时间都控制在 15 ~ 20min。

最后，测试人员中要有音频专家，他们知道该听什么，以及该怎么听，这个可以保证测试的可重复性。但也不能都是专家，大部分应该是普通人。这是因为普通人的意见更有普适性。他们对一些明显的缺陷比如“颤音”、“漏音”比较敏感，而对一些声音的细微差异可能会给出相近的分数，然而这样更符合大众的听音习惯。如果都是专家则分数过于严苛反而偏离了人群总体的听感。

这里举一个用 MUSHRA 给不同语音合成算法打分的例子，具体如图 1 所示。图中的 REF 就是无损的音源，Anchor35 就是我们说的锚点（预估只有 35 分）。这里还把无损音频的采样位深调为 8bit，标记为 8bit μ-law REF。而其它则是代表不同的语音合成算法。

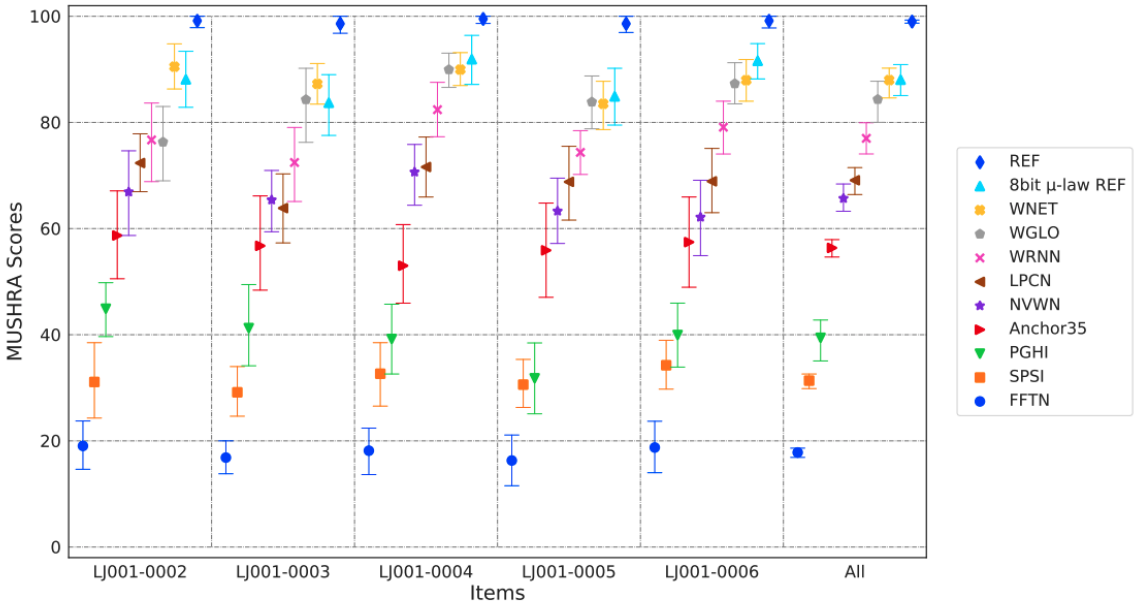


图1 不同语音合成方法的MUSHRA评分（Govalkar, Prachi, et al.）

我们可以看到 WNET 的分数最高很接近无损音源，这代表这种算法的语音生成质量最好。锚点一般是质量很差的音频，在图 1 中甚至有些算法的分数比锚点还低，这代表他们的生成效果极差。这些极差的算法甚至还抬高了锚点的分数，我们可以看到 Anchor35 最后的 MUSHRA 评分在 50～60 分之间。

为了方便人们测试使用，GitHub 上有基于Web 版的MUSHRA 自动打分工具，你可以自行下载使用。

常用的主观评价标准

MUSHRA 方法的测试面比较广，可以用于编 / 解码器，语音合成，甚至是耳机测评。除了 MUSHRA，**在 ITU 中还有其它一些针对不同场景的音频测评标准**，比如 ITU-T 评价标准中的 ITU-T P.800《语音质量的主观评价方法》，也就是我们常说的 MOS（Mean Option Scores，平均意见分），以及 ITU-T P.830《电话和宽带数字语音编码器的主观评价方法》、ITU-T P.805《对话质量的主观评价》等。ITU-R 主观评价标准中的 ITU-R BS.1116 《音频系统中小损伤主观评价方法》、ITU-R BS.1285 《音频系统中小损伤主观评价的预选方法》等。如果你有兴趣可以上ITU 官网自行查看。

音频客观评价方法

主观评价的缺点是：人少了、执行不规范都会带来测试偏差。那么有没有什么客观评价的方法，可以给出可复现又贴近主观评价的结果呢？下面我就来为你介绍一下符合这些要求

的客观评价方法。

客观评价主要包括有参考评价和无参考评价。所谓有参考评价就是除了测试音频以外，还需要同时给出一个参考音频做为基准，通过计算测试音频和参考音频的区别来拟合出音频的主观得分。而无参考的客观评价则不需要参考音频，直接根据音频的频谱能量分布、连续性等指标来评分。

有参考音频质量评价

我们先来看一下有参考的客观评价方法。在 2001 年，ITU-T P.862 标准定义了有参考客观评价算法 PESQ（Perceptual Evaluation of Speech Quality，语音质量感知评价），该算法主要用来**评估窄带（8kHz 采样率）及宽带（16kHz 采样率）下的编解码损伤**。该算法在过去的二十年中，被广泛的应用于通信质量的评定。

随着技术的发展，PESQ 的应用范围变得越来越窄，于是在 2011 年，P.863 标准定义了一套更全面、更准确的有参考客观评价算法 POLQA。相比 PESQ，**POLQA 可评估的带宽更广，对噪声信号和延时的鲁棒性更好，其语音质量评分也更接近于主观的评分**。

PESQ 算法已经开源，而 POLQA 你需要购买一套专门的设备和授权才能使用。所以目前做一些日常的测试中，PESQ 还是用的比较多的方法。这里我们主要介绍一下 **PESQ 算法的基本原理**。PESQ 算法的处理步骤如图 2 所示：

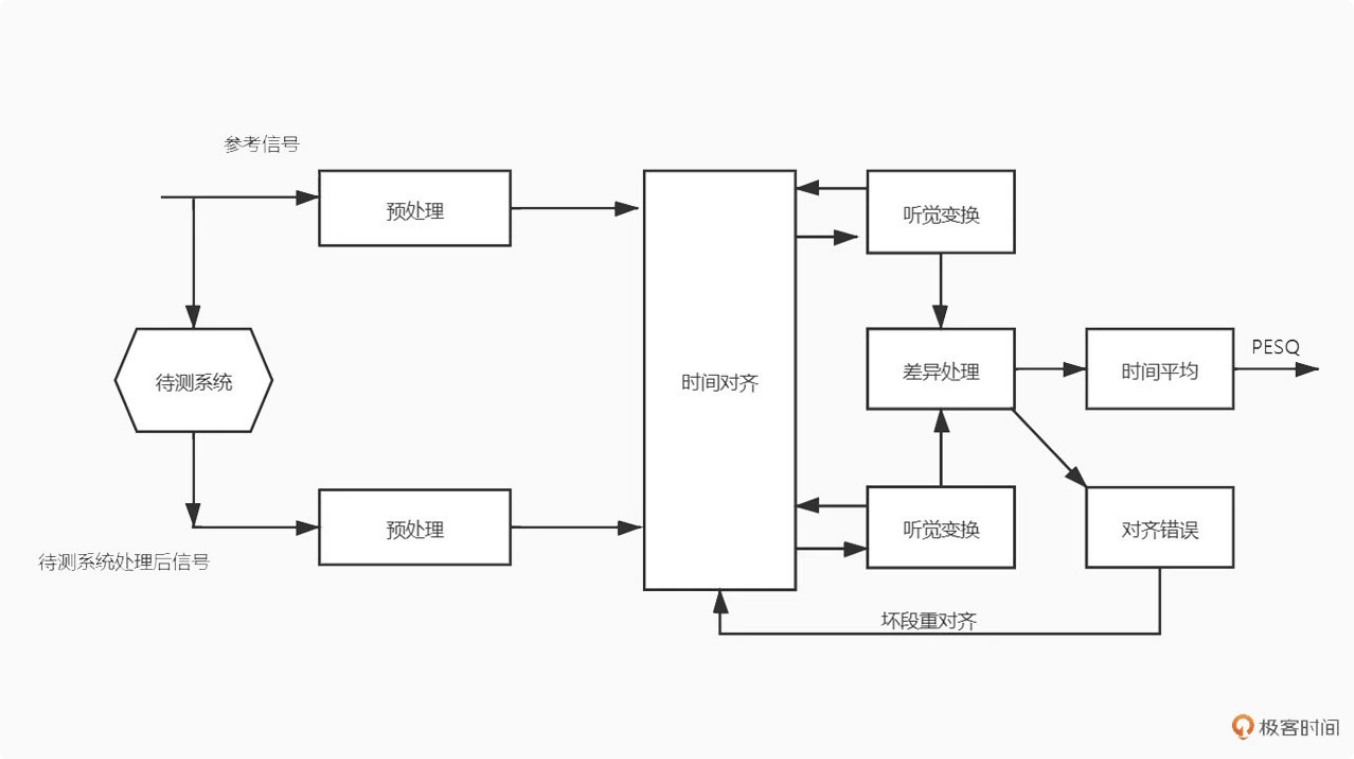


图2 PESQ处理步骤

图 2 中待测系统就是你的音频系统或者算法，比如一个编 / 解码器。为了消除系统延迟的影响，首先，将参考信号和系统处理后的信号经过相同的预处理后进行时间对齐；然后，进行听觉变换把音频信号转化为频谱信号；接着，再对能量谱逐帧进行差异处理；最后，取时间平均得到 PESQ 分数。如果发现有的音频片段差异特别巨大，则表明存在对齐错误，需要对没对齐的片段进行再对齐。

PESQ 的分数范围在 0～4.5 分，一般音质比较好的编 / 解码器，比如 64kbps 比特率的 OPUS 编解码器，可以达到 4.5 分，而分数越低则代表音质越差。比如 OPUS 的码率降到 6kbps 那 PESQ 可能就只有不到 3 的分数。

值得注意的是，PESQ 最多只能评价 16kHz 采样率的音频。如果要评价一个采样率比较高的音频信号，比如音乐信号，POLQA 会比较合适。POLQA 最高可以支持 48kHz 采样率的全带音频的客观质量评价。购买一套 POLQA 设备的价格都是百万级的，为了方便使用，你还可以考虑一下使用例如 ViSQOL 等开源算法，也可以支持 48kHz 的音频采样率。

无参考音频质量评价

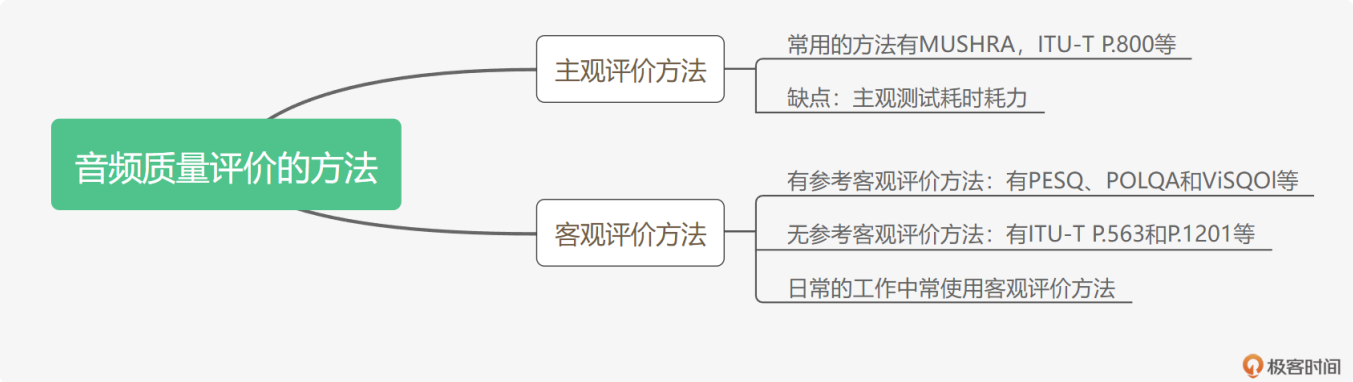
有的时候我们可能无法获得参考音频，比如在打网络电话时，只有接收到的经过编 / 解码和网络传输的音频信号，**没有远端的输入信号**。这时候无参考音频质量评价方法就派上用场了。**不需要参考信号，仅通过对输入信号本身或参数的分析即可得到一个质量评分**。比较著名的无参考客观评价方法有 ITU-T P.563、ANIQUE+、E-model、ITU-T P.1201 等。

其中，ITU-T P.563 于 2004 年提出，主要是面向窄带语音的质量评估；ANIQUE+ 于 2006 年提出，也是面向窄带语音，其评分准确度据作者称超过了 PESQ，不过 PESQ 的测量不能反应网络的延时、丢包等，并不能完美适用于如今基于互联网传输的实时互动场景。E-model 于 2003 年提出，不同于上述两种方法，这是一个基于 VoIP 链路参数的损伤定量标准，不会直接基于信号域进行分析。ITU-T P.1201 系列于 2012 年提出，对于音频部分，该标准也不对音频信号直接进行分析，而是基于网络状态和信号状态对通信质量进行评分。

其实在实际使用中由于实时音频处理过程复杂，除了编解码器和网络对音频可能造成损伤，音频的处理步骤比如降噪、回声消除等步骤，也可能对音频的质量造成影响。现有的无参考音频质量评价还不能准确地反映音频的实际质量情况，或者说使用的时候还有很多限制，比如隐私问题无法获得音频信号或者音频链路的信息指标无法准确获得。所以无参考音频质量评价还有很多有待研究的地方。

小结

好了，为了方便你记忆，这里我用图来总结一下音频质量评价的方法。



注意客观测试目前只能测量音频的损伤程度，比如我们平时在选择编解码器时不知道选择什么类型、多少码率，这时不妨跑一下 PESQ 或者 VisQol 看一下分数。而对于一些偏主观的测试类型，比如音乐听感、耳机效果等，则一般还是需要依靠主观测试。主观测试有时候也不需要你自己组织，你也可以采用众包等方式把测试分发出去。针对这种分发式的主观测试，最近的 ITU-T P.808 也有详细的测试流程和方法，如果你有兴趣，可以自行了解一下。

思考题

通过这节课的学习，你掌握了多少知识呢？我这里有一道思考题留给你。

如果我们想对一款在线通话 App 做实时通话质量打分，我们可以采用什么样的音频评价方法？

你可以把你的答案和感受写下来，分享到留言区，与我一起讨论。我们下节课再见。

分享给需要的人，Ta订阅后你可得 20 元现金奖励

分享给需要的人，Ta购后你可得 20 元现金奖励

生成海报并分享

赞 0 提建议

© 版权归极客邦科技所有，未经许可不得传播售卖。 页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 03 | 如何分析与处理音乐信号？

下一篇 05 | 音频降噪如何对症下药？

精选留言 (5)

写留言

晓龙
2021-11-30

可以用主观法：
1. 通话时，分别在发送端和接收端做录音，然后对比。
2. 语音延迟，可以用双端报数的方式主观测试。

无参考客观法，可以通过查看音频丢包率，网络延时等评测

展开

作者回复: 嗯是的，稍微补充一下后续的课程里可能会发现，降噪和回声消除也会对主观分数产生影响。

写点啥呢
2021-11-29

请教老师，无参考客观评价方法，其评价算法的设计思路和指标设计是什么样子的，比如它考虑哪些指标来给出评价呢？

展开

作者回复: 这个不同算法可能不太一样。举个例子，每个中间模块的状态。比如网络延迟、抖动，codec自身的MOS、以及一些音频前处理中的信噪比例、信回比例之类的指标来综合评价。

徐刚

2021-11-29

客观评价无参考方法；能够实时反馈通话质量；可以结合ANIQUE+ + ITU-T P.1201；实时网络通话，包含前段处理和网络传输，这2者结合，能够覆盖整个通话链路，反应最终通信双方实际的听觉感受

作者回复: 正确

Geek_7de4c5

2021-11-29

P.563和P.1201同时打分，再算均分比较合适

展开

springXu

2021-11-29

在线实时通信的音频测试，我认为用客观无参考测试的方案中去选。原因是网络好坏也会改变通话质量，这个参考音频是没有的。 但对于知道了音频的帧率和编码解码时，有了参考样本，可选的测试范围就多了。

展开

作者回复: 回答得挺对